Journal of Memory and Language 117 (2021) 104189

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

The dissociations of confidence from accuracy in forced-choice recognition judgments

Maciej Hanczakowski [a,*], Ewa Butowska [a], C. Philip Beaman [b], Dylan M. Jones [c], Katarzyna Zawadzka [a,*]

[a] Interdisciplinary Center for Applied Cognitive Studies, SWPS University, Poland
[b] School of Psychology and Clinical Language Sciences, University of Reading, UK
[c] School of Psychology, Cardiff University, UK

ARTICLE INFO

Keywords:
Confidence
Metacognition
Forced-choice recognition
Calibration

ABSTRACT

Judgment of confidence in memory is likely to track memory accuracy if those factors shaping accuracy also shape confidence. In recognition memory, accuracy is determined by the relative level of evidence present for the target and that supporting the lures. As the discrepancy between targets and lures increases, so does the likelihood of correct responding. In contrast, this study shows that confidence can instead depend on the *absolute* evidence supporting the chosen target rather than the balance of evidence between targets and lures. In four experiments, using different types of forced-choice recognition tests, we demonstrate that generally manipulating the strength of evidence supporting targets affects confidence judgments but that varying the strength of evidence supporting lures creates robust confidence-accuracy dissociations, changing accuracy while not affecting confidence. Together, these data support an absolute account of confidence in forced-choice recognition and demonstrate that confidence-accuracy dissociations across recognition conditions are likely to be ubiquitous.

Introduction

Everyday disclosures involving memory are often accompanied by statements of confidence. Thus, for example, a witness trying to describe a person who committed a crime might say that 'He wore a black jacket, I am 100% certain of this'. Such expressions of confidence are the topic of studies concerned with establishing how confidence judgments relate to memory performance. This can be assessed either for confidence in future memory performance, expressed in so-called judgments of learning (Dunlosky & Nelson, 1992; Koriat, 1997; Zawadzka & Higham, 2016), or confidence in past memory reports, often referred to as retrospective confidence judgments (Mazancieux, Fleming, Souchay, & Moulin, 2020; Pleskac & Busemeyer, 2010). There are numerous examples in the literature where conditions giving rise to higher judgments of learning actually result in memory performance that is no different or even lower than the comparison conditions (e.g., Besken & Mulligan, 2013; Rhodes & Castel, 2008; Undorf, Zimdahl, & Bernstein, 2017), but this is less often reported for retrospective confidence judgments. In the present study, we examine a specific format of memory testing in which participants have to choose among two potential answers to a memory question – the two-alternative forced-choice recognition test – and scrutinize conditions under which retrospective confidence judgments become dissociated from memory performance.[1]

The issue of retrospective confidence judgments is particularly important when the outcome of the memory test has material consequences, such as for eyewitness testimony. A recent surge in discussion of confidence judgments in the context of memory reports is related to the issue of line-ups. Line-up identification comprises a more complex version of a forced-choice recognition test, with an additional option of rejecting a line-up if the culprit is not included. Investigators and jurors often rely on expressions of confidence accompanying identification decisions in such situations (see Wixted, Mickes, Clark, Gronlund, & Roediger, 2015, for a discussion). In this context, it is vital to know whether high confidence can be interpreted in terms of high likelihood that a corresponding memory judgment is accurate. There is general agreement that, at least for eyewitnesses who choose a suspect from a

* Corresponding authors at: Interdisciplinary Center for Applied Cognitive Studies, SWPS University, Chodakowska 19/31, 03-815 Warszawa, Poland.
E-mail addresses: maciej.hanczakowski@gmail.com (M. Hanczakowski), k.n.zawadzka@gmail.com (K. Zawadzka).

[1] Given that our study is concerned solely with retrospective confidence judgments, throughout the paper the word 'confidence' should be interpreted as pertaining to retrospective confidence.

https://doi.org/10.1016/j.jml.2020.104189
Received 12 February 2020; Received in revised form 29 October 2020; Accepted 7 November 2020
Available online 20 November 2020

line-up, identification confidence typically tracks accuracy (e.g., Brewer & Wells, 2016; Wixted & Wells, 2017), at least when memory is tested for the first time. Thus, identifications made with the highest levels of confidence can be generally considered to be also highly accurate (but see Grabman, Dobolyi, Berelovich, & Dodson, 2019, for exceptions). Similarly, in experimental studies on memory and confidence, using more artificial materials like pictures or lists of words, the positive relationship between the two is well documented, inasmuch as more accurate responses are accompanied by higher confidence, at least for materials that are not deceptive (Brewer & Sampaio, 2006; Koriat, 2012). While this overall empirical relationship is well established, there is relatively little work to show *how* confidence judgments are formed in the memory domain, or why, in some circumstances, they fail to track accuracy.

Confidence-accuracy relationships can be analyzed in a range of ways, delineated in a thorough overview by Roediger, Wixted, and DeSoto (2012; see also Busey, Tunnicliff, Loftus, & Loftus, 2001). One question, for example, may be whether on average responses that a given person provides with high confidence are more likely to be correct than responses which the same person provides with lower confidence. This approach – referred to as an investigation of metacognitive resolution – is often vital to studies concerned with self-regulated remembering (e.g., Higham, 2002, 2007; Koriat & Goldsmith, 1996), as well as eyewitness research, but it does not allow for a systematic analysis of the bases of confidence-accuracy relationships. For this purpose, an experimental approach is necessary where independent variables are manipulated and their influence on accuracy and confidence examined. In other words, when the experimental approach is adopted it is not the influence of experimental variables on resolution that is of interest but rather the influence of these variables separately on accuracy and confidence. The confidence-accuracy relationship is then derived from assessing whether a given manipulation affects both average accuracy and confidence in the same way – a positive relationship across experimental conditions – or whether a given manipulation dissociates confidence from accuracy, when a positive relationship fails to emerge.

The experimental approach to confidence-accuracy relationships can take two forms. Researchers can find a variable which affects confidence and examine whether it exerts the same type of influence on accuracy, or they can find a variable which affects accuracy and examine whether it exerts the same type of influence on confidence. There is ample evidence for cases in which confidence is shaped by factors that do not affect accuracy. For example, positive feedback after a recognition decision is highly likely to increase one's confidence in this decision while obviously no longer being able to affect its accuracy (Semmler, Brewer, & Wells, 2004). Also, spurious familiarity of memory items may affect confidence in recognition decisions when the accuracy of these decisions depends on associative information rather than item familiarity (Hanczakowski, Pasek, Zawadzka, & Mazzoni, 2013). Thus, confidence-accuracy dissociations can be created when factors that do not exert influence on accuracy of memory responses are, nevertheless, used to arrive at confidence assessments for these responses.

The present study, on the other hand, focuses on the second, less often adopted form of the experimental approach to investigating confidence-accuracy relationships. Here we ask the question of how confidence is affected by experimental manipulations known to affect memory accuracy in forced-choice recognition tests. This is a vital problem inasmuch as it speaks to the prevalence of confidence-accuracy dissociations under those testing conditions. If every single manipulation that affects memory accuracy can be expected to affect confidence in the same way, then confidence should track accuracy across a wide variety of encoding and retrieval conditions. The overall level of confidence can then be generally used to infer the quality of encoding and retrieval unless a particular misleading cue can be identified that is likely to spuriously affect confidence.

The received view on whether factors affecting accuracy generally affect confidence in the same way is that they almost uniformly do. For example, Pleskac and Busemeyer (2010) in their overview of confidence models identified – based on numerous studies from various domains – the pattern of 'positive relationship between stimulus discriminability and observed confidence' (p. 869) as one of the main empirical regularities that any viable model of confidence needs to explain. Importantly, the same has been claimed in reference specifically to studies on memory reporting. In their overview of research on confidence-accuracy relationship in recognition, Roediger et al. (2012) concluded that "confidence and accuracy seem well correlated in this kind of experiment in which independent variables are manipulated. In fact, the exceptions are sufficiently few that we can safely conclude that when an independent variable affects accuracy of memory reports, subjects' confidence in those reports will virtually always be affected the same way (however, see Tulving, 1981, for a somewhat different case)" (p. 98).

Before moving to the discussion of a study by Tulving (1981) – one of seemingly very few exceptions to a generally strong confidence-accuracy relationship in experimental research (see Busey et al., 2001; Chandler, 1994; Starns & Ksander, 2016, for other examples) – it is important to consider the reasons why such a generally strong relationship might exist. One possibility is that people have privileged access to the contents of their own memory, allowing them to be confident when they feel that the right information is stored in memory. This gives rise almost inevitably to correct responses and results in lower confidence only when relevant information is not stored, a condition which also produces more erroneous responses. However, despite its intuitive appeal, this so-called direct-access view (Hart, 1965) has long been abandoned in research on metacognitive judgments – of which retrospective confidence judgments are an example – in favor of a cue-utilization view (Koriat, 1997).

According to the cue-utilization view, people formulate their metacognitive judgments based on a variety of cues that are often by-products of the process of memory search itself. Thus, for example, people are confident in their responses when relevant memories are retrieved fluently (Kelley & Lindsay, 1993). Since correct memories are also, by and large, retrieved more fluently than erroneous responses, this generally results in a positive confidence-accuracy relationship. Importantly, a manipulation that positively affects fluency of retrieval – for example greater opportunity for rehearsal of to-be-remembered items (Busey et al., 2001) – will simultaneously benefit memory performance and increase confidence. Thus, positive confidence-accuracy relationships across experimental conditions are to be expected when confidence is based on the same cues that determine the accuracy of recognition decisions (Reinitz, Peria, Séguin, & Loftus, 2011). When these cues are different – such as in the case of feedback that follows recognition decisions – confidence and accuracy start to diverge.

To achieve complete understanding of the confidence-accuracy relationship in recognition decisions one thus needs to consider more fully what determines recognition accuracy. This question is vital as it enables assessing whether the same factors influence confidence. On one level, this is a daunting task with which researchers have been struggling for many years. However, at the very basic level of analysis it is reasonable to suppose that the accuracy of forced-choice recognition decisions depends on balance of memory evidence supporting alternatives in this test. Memory evidence might reflect a continuous unitary signal (Wixted, 2007), a conglomerate of different signals (Yonelinas, 1994, 2002), or discrete memory states (Bröder & Schütz, 2009), but no matter which conceptualization is adopted, it remains the case that the stronger the evidence gathered in support of targets and the weaker the evidence gathered in support of lures, the more accurate the responding in this test will be. This holds even if a person does not consider both alternatives. Recently, it has been argued that in forced-choice recognition tests participants can turn the decision on some trials into a quasi-old/new recognition test (Jou, Flores, Cortes, & Leka, 2016; Starns, Chen, & Staub, 2017). This involves considering only one alternative and making an absolute judgment, endorsing it if it crosses a certain

criterial level of evidence. Even in this situation, stronger evidence for targets and weaker evidence for non-targets should shape correct responding.

If accuracy in forced-choice recognition depends on a difference in evidence supporting recognition alternatives, then one can expect a strong confidence-accuracy relationship if confidence is also based on this difference. If it is, then any change in the balance of evidence will be reflected in both accuracy and confidence, giving rise to a positive confidence-accuracy relationship across experimental conditions. The fact that positive accuracy-confidence relationships are ubiquitously observed – as summarized by Roediger et al. (2012) – has resulted in models of retrospective confidence judgments in forced-choice tests which explicitly assume that confidence is indeed based on the balance of evidence between alternatives (Horry & Brewer, 2016; Pleskac & Busemeyer, 2010). However, although such relative accounts of confidence continue to be widely held, recent lines of evidence suggest that confidence in forced-choice decisions generally, and forced-choice recognition in particular, may not necessarily be based on the balance of evidence.

Empirical patterns which point to confidence in forced-choice decisions reflecting solely evidence supporting the chosen alternative – what will be henceforth referred to as the *absolute account* of confidence in forced-choice decisions – first emerged in research on perception. Zylberberg, Barttfeld, and Sigman (2012) examined retrospective confidence regarding decisions concerning luminance of stimuli or direction of movements and revealed that confidence was independent of evidence supporting the unchosen alternative, contrary to the predictions of the relative, balance-of-evidence account. More importantly from the present perspective, analogous results have been described in the domain of memory. Jou et al. (2016) were the first to suggest that in forced-choice recognition tests participants may also base their confidence on the assessment of evidence supporting only the chosen alternative. In their Experiment 1B, Jou et al. administered a standard two-alternative forced-choice recognition test for studied words, coupled with confidence judgments. Somewhat untypically, some trials on this test consisted of two studied words as alternatives. Surprisingly, confidence for such deceptive trials was higher than for the more conventional trials in which one studied word and one non-studied word were presented as recognition alternatives. This pattern of results is not consistent with the balance-of-evidence account of confidence inasmuch as the difference in evidence should be larger on trials consisting of a studied and a non-studied alternative (target-lure pair) than on those with two studied alternatives (target-target pair). However, this pattern can be accommodated by the absolute account of confidence, in which an alternative chosen from a target-target pair – essentially the stronger of two studied words – should be associated with more evidence than an alternative chosen from a target-lure pair, where the target needs to exceed a lower criterion of strength in order to be endorsed.

That confidence in forced-choice recognition decisions depends on absolute evidence for the chosen alternative – rather than the balance of evidence across alternatives – has recently been confirmed in a study by Zawadzka, Higham, and Hanczakowski (2017), who examined confidence judgments in forced-choice recognition using the plurals paradigm (Hintzman, Curran, & Oppy, 1992). The use of the plurals paradigm allowed the strength of targets and lures to be varied independently via the number of presentations of their parent words (e.g., 'frogs' constituted a strong lure if 'frog' was studied multiple times). This study showed that, when making confidence judgments, participants considered evidence supporting the chosen alternative while at the same time virtually ignoring the strength of evidence supporting the unchosen alternative, thus creating a dissociation between accuracy and confidence: Participants were more likely to choose a target when a lure was weak but still confidence reflected only support for the target and not support for the lure.

Further evidence for the absolute account of confidence is provided by Miyoshi, Kuwahara, and Kawaguchi (2018), who focused on forced-choice recognition decisions concerning pictures. They demonstrated that when incorrect responses are made for recognition trials including highly memorable studied pictures, confidence is higher than when incorrect responses are made on trials including studied pictures of lower memorability. Once again, this pattern is the opposite of what the balance-of-evidence account would predict but it is easily accommodated within the absolute account if the endorsed incorrect alternatives are particularly strong when pitted against highly memorable competitors.

The studies by Jou et al. (2016), Zawadzka et al. (2017), and Miyoshi et al. (2018) demonstrate that confidence in forced-choice recognition may *not* depend on the difference in evidence supporting alternatives at test, which is the factor that shapes the accuracy of recognition decisions. Hence, confidence-accuracy dissociations across experimental conditions should be observable when the strength of evidence supporting lures is varied. Such manipulations should change the balance of evidence across alternatives, yet at the same time not affect confidence (in correct responses), which seems to depend on the strength of endorsed targets. The significance of such dissociations has passed unnoticed until now, possibly because studies concerned with forced-choice recognition have employed manipulations that systematically varied only the strength of targets (e.g., Palmer, Brewer, Weber, & Nagesh, 2013). Consistent with this idea, the one demonstration of a confidence-accuracy dissociation across experimental conditions cited earlier – the study of Tulving (1981), to which we now return – included manipulations of lures rather than targets.

In Tulving's study (1981; see also Dobbins, Kroll, & Liu, 1998; Heathcote, Bora, & Freeman, 2010, for conceptual replications), participants studied halves of exterior scenes that included landscapes, buildings, etc. In a subsequent forced-choice recognition test, three conditions were included. In the A-A' condition, old halves (A) were paired with their corresponding non-studied halves (A'). In the A-B' condition, old halves were paired with non-studied halves of different studied images (B'). In the A-X condition, old halves were paired with novel halves of pictures, not corresponding to any of the studied pictures (X). With these conditions, Tulving was able to bring about two confidence-accuracy dissociations.

In the first dissociation, while performance in the A-A' condition was slightly higher than performance in the A-B' condition (a difference of six percentage points in Experiment 1), confidence in correct responses was actually higher in the A-B' condition. This can be explained by assuming that both accuracy and confidence depend on the difference in evidence supporting the alternatives, with an additional important contribution of variance of this difference, which is restricted in the A-A' condition (because evidence for both alternatives is underpinned by the same memory representation) but increased in the A-B' condition (see Clark, 1997). Limited variance of difference in evidence means that even small differences in evidence are highly diagnostic and this serves to boost recognition accuracy, while small average differences in evidence between two alternatives reduce confidence. We will return to this issue later, in Experiment 4 of the present study.

The second dissociation documented by Tulving (1981) – which to the best of our knowledge has not attracted attention of other researchers – is that that there was no appreciable difference in confidence in correct responses between the A-B' condition and the A-X condition, while at the same time performance in the A-X condition was markedly higher than in the A-B' condition (a difference of 19 percentage points in Experiment 1). This striking pattern can be explained if one assumes that confidence does not rely on the difference in strength of alternatives but instead is dependent solely upon the strength of the chosen alternative. When comparing A-B' to A-X this strength of the chosen target should be similar, determined mostly by the memory trace of A, hence confidence in correct decisions should be roughly equivalent. However, the task is more difficult when pitting two alternatives which are each similar to studied items (as in the A-B' condition) than when pitting one old against one completely novel alternative against each other (as in the A-

X condition). This results in equivalent confidence judgments, along with a marked difference in recognition performance: a confidence-accuracy dissociation.

The novel account of confidence in recognition decisions put forward here suggests that confidence in forced-choice recognition judgments is subject to a form of confirmation bias (Nickerson, 1998), in that it assumes that confidence depends on absolute evidence supporting the chosen alternative rather than on relative evidence for all recognition alternatives. In doing so, it explains the results obtained by Tulving (1981) for the A-B' and A-X comparison and also suggests that confidence-accuracy dissociations in forced-choice tasks may be much more prevalent than previously thought. In a nutshell, a dissociation should be observed whenever forced-choice recognition performance is reduced by making lures stronger, thus modulating the difference in evidence between target and lure, but holding the absolute evidence for the target (and hence confidence in the choice) constant.

This principle of absolute evidence also sheds new light on previous results from our own group. Beaman, Hanczakowski, and Jones (2014) examined the role of auditory distraction on metacognitive monitoring. Only a subset of these results is relevant here, the inadvertent discovery of a striking confidence-accuracy dissociation: a difference in accuracy across conditions was accompanied by a difference in confidence in the opposite direction. This dissociation was obtained – though not discussed at the time – when accuracy and confidence responses were recorded across two types of recognition tests for lists of randomly paired nouns. In an *associative recognition test*, participants were presented with one pair intact from the study phase and one pair comprising study words in a novel combination. The task here was to distinguish the familiar pairing from the unfamiliar pairing. The *item recognition test* used the same study items, but the two pairs of items presented at test now comprised one pair of items rearranged from study and one pair comprising two novel words. The task here was to identify the pair made up of the two studied words, regardless of their original pairing[2]. Accuracy in the item recognition test was higher than in the associative recognition test, but at the same time confidence was lower.

These results can be readily explained by the absolute account of confidence in forced-choice recognition. Unsurprisingly, because the item recognition test used novel words as lures, the difference in strength of memory evidence between targets and lures is greater in this test than in the associative test that comprises words already encountered in the experiment. Accordingly, accuracy is greater in the item recognition test. A different pattern emerges when confidence is considered. The target pair in the associative recognition test is supported by evidence for both individual words and their association. By contrast, the basis for choice of the target pair in the item recognition test rests only on the evidence for individual words: The pairing of words is not germane to recognition accuracy. If confidence depends solely on evidence supporting the chosen alternative—an absolute judgement that ignores the evidence for the unchosen alternative—then it should be paradoxically higher in the more difficult associative recognition test, as observed by Beaman et al. If these results reflect the workings of an absolute evidence heuristic of yielding confidence judgments, then this significantly broadens the remit of this particular account of confidence in forced-choice recognition by demonstrating its workings within complex memory tasks involving both item and associative information. While previous studies that can either be interpreted as consistent with the absolute account (Jou et al., 2016; Tulving 1981) or were designed to specifically test this hypothesis (Miyoshi et al., 2018; Zawadzka et al., 2017) used simple materials such as individual words or pictures, the study by Beaman et al. (2014) required choosing among pairs of words that differed both in item information and associative information. The observed crossover confidence-accuracy dissociation suggests that both in tests that require access to item information (the item recognition test) and tests that require access to associative information (the associative recognition test), the same principles for rendering confidence judgments are likely to operate.

However, while suggestive, the Beaman et al. (2014) results are not definitive inasmuch as they were an unanticipated outcome of a setting designed for other purposes. Additionally, one technical feature of their procedure is troubling: the length of the test list between associative and item recognition tests was equated, and equal to the number of the pairs in the study list. This meant that for the item recognition test each word was used once during the test but for the associative recognition test each word was actually used twice – once for the intact pair and once for a rearranged pair. Thus, individual words accrued more familiarity in the associative recognition test by virtue of their repeated presentation in the test phase. The pattern of confidence documented by Beaman et al. might therefore be explicable by this increased familiarity with test words in the associative recognition test rather than by the absolute account of confidence judgments in forced-choice recognition.

The current study presents four new experiments free of this confound and demonstrate how confidence in forced-choice recognition is a function of memory evidence supporting the chosen alternative, leading to robust dissociations of confidence and accuracy of recognition decisions. We first isolate the dissociation, then provide a direct test of the absolute account of confidence-accuracy relationships that attributes accuracy to the difference in strength between the assessed alternatives and confidence solely to the strength of chosen alternatives. Finally, we broaden the scope of our investigation by demonstrating how confidence in forced-choice recognition can still be sensitive to the type of lures used in particular test trials, even when evidence supporting these lures is not factored into confidence judgments. Together, these experiments show that, decades of theorizing notwithstanding (see Pleskac & Busemeyer, 2010, for a review), changes in forced-choice recognition accuracy across experimental conditions need not be associated with parallel changes in confidence.

Experiment 1

Here, we compare accuracy and confidence levels in an associative recognition test and an item recognition test that used rearranged pairs as targets. A schematic outline of the two types of test used in this experiment is presented in Fig. 1. In the associative recognition test, on each trial an intact studied pair was presented along with a pair of words drawn from different studied pairs, and the task was to pick the pair that was presented in the study phase. Here, all words were equally familiar – the item information is equated – but the target pair also had the association created at study. Accordingly, in the associative recognition test alternatives differ in associative information only. In the item recognition test, rearranged pairs served as targets, and lures were created by pairing two novel words, not presented before in the experiment. In the item recognition test alternatives differ mostly with respect to item information, as even the target has no associative information as legacy from the study phase. Participants' task here was to pick the target containing words from study, necessarily drawn from two different pairs in the study list.

Based on the results presented by Beaman et al. (2014), we expected greater accuracy in the item recognition test which used familiar words – albeit in a novel pairing – as targets, and novel, unfamiliar words as lures, in contrast to the associative recognition test with familiar targets and familiar lures. The main focus of the experiment is on confidence patterns. We expected to replicate the pattern of lower confidence levels for correct responses in the item recognition test. This pattern of lower confidence in correct item recognition decisions, despite higher overall accuracy, was predicted because confidence reflects support for the chosen alternative only, not the difference in evidence between alternatives. Targets should be stronger in the associative recognition test,

[2] Beaman et al. (2014) used different names for the tests, referring to the item recognition test as the recombined test, and referring to the item+association test, introduced here in Experiment 3, as the simple recognition test.

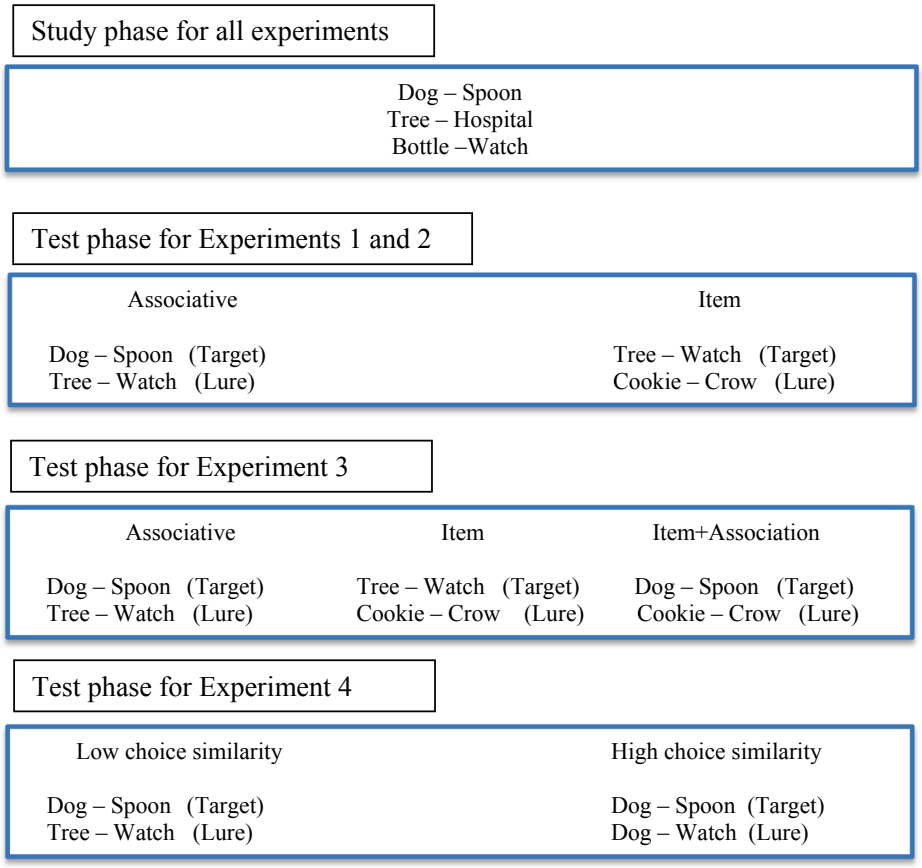

Fig. 1. The examples of word pairs used as targets and lures in different types of test conditions in Experiments 1–4. All experiments used unrelated word pairs as study materials. Experiments 1 and 2 contrasted two testing conditions: (1) Associative, where intact pairs served as targets and rearranged pairs served as lures, (2) Item, where rearranged pairs served as targets and novel pairs served as lures. Experiment 3 added a novel testing condition – Item + Association – where intact pairs served as targets and novel pairs served as lures. Thus, the Item + Association condition had the same targets as the Associative condition and the same lures as the Item condition. Experiment 4 used only associative recognition tests and contrasted conditions of Low choice similarity, with intact and rearranged pairs composed of different words, and High choice similarity, with intact and rearranged pairs sharing the first word.

since target memory strength is a joint function of both item and associative information. In the item recognition test memory for the association would be absent, giving lower memory strength leading to lower confidence.

Method

Participants

Forty Cardiff University students participated in exchange for course credit. The sample size was chosen to be comparable to the one used by Beaman et al. (2014), who tested 42 participants.

Design

The experiment used a 2 (test type: associative recognition vs. item recognition) × 2 (test length: short vs. long) design. The conditions were manipulated within participants but between studied lists, with one list of pairs of words assigned to each of the four conditions. This assignment was counterbalanced across participants.

In the associative test, participants were asked to distinguish between intact and rearranged pairs, whereas in the item test participants were asked to distinguish between rearranged and novel pairs. The test length factor manipulated the effect of the number of presentations of individual words at test. Specifically, the *long test condition* included the confound present in the study by Beaman et al. (2014) with each studied pair from a list assigned to the associative test serving as a source of words for one intact and one rearranged test pair. In this way, each individual word was presented twice. In the item test, each studied word was included in only one rearranged pair and thus was presented at test once only. By contrast, the *short test condition* removed this confound. Thus, each studied pair from a list assigned to the associative test condition served either as a source of words for an intact or a rearranged pair. This meant that only half of a studied list could be used for creating targets in this test, so that the length of the test was half that experienced in the long test condition. A short item test, also using only half of studied pairs as a source for test pairs, was included in the design to control for the test length (thus equating fatigue, output interference, and so on, in both short test conditions).

Materials

A cohort of 440 words was chosen from the MRC database of which 320 were used to create four lists of 40 pairs of words for study and the remaining 120 words were used to create novel pairs for the item recognition tests. For each list, four types of test were created.

For the long associative test, all pairs were included as 40 intact pairs and words from all pairs were also reshuffled to create 40 rearranged pairs. For the short associative test, half of the pairs were randomly chosen to serve as 20 intact pairs and the words from the other half were recombined to create 20 rearranged pairs. For the long item test, words from all pairs were recombined to create 40 rearranged pairs and additional 40 pairs of unstudied words were added as novel pairs. For the short item test, half of the pairs were randomly chosen and the words from these pairs were recombined to create 20 rearranged pairs; an additional 20 pairs of unstudied words were also added as novel pairs.

Procedure

Participants were tested on individual computers. They were given four lists of 40 pairs to study—each pair shown for 1500 ms, with a 500 ms interstimulus interval—and each list was immediately followed by a self-paced test in one of the four experimental conditions. The order of conditions was randomly determined for each participant.

The nature of the target and lure pairs for each test was carefully explained in the instructions preceding the test. The tests took form of two-alternative forced choice (2AFC) recognition. Target pairs (either intact or rearranged) were presented simultaneously with a lure pair

(either rearranged or novel). For each test trial, participants were asked to endorse the target pair. Immediately after providing their response, participants were asked to judge their confidence that this response was correct by typing in their confidence judgment on a scale from 50 (chance level in a 2AFC test) to 100%.

Results and discussion

The descriptive statistics are given in Table 1. Hit rates were subjected to a 2 (test type) × 2 (test length) repeated-measures ANOVA, which yielded a significant main effect of test type, $F(1, 39) = 7.96$, $MSE = .04$, $p = .007$, $\eta_p^2 = .17$, with accuracy higher in the item ($M = .81$, $SD = .15$) than in the associative test ($M = .72$, $SD = .15$), but no main effect of test length, $F(1, 39) = 0.15$, $MSE = .01$, $p = .696$, $\eta_p^2 = .004$, and no interaction, $F(1, 39) = 1.27$, $MSE = .01$, $p = .267$, $\eta_p^2 = .03$. This confirms higher performance in the item recognition test.

The analyses of confidence data focused on confidence in correct responses, for which the absolute account of confidence makes specific predictions. Confidence in incorrect decisions was not analyzed because generally high levels of memory performance resulted in noisy data, with many incomplete cells. Here, the main effect of test type, $F(1, 39) = 4.97$, $MSE = .01$, $p = .032$, $\eta_p^2 = .11$, stemmed from significantly *lower* confidence in the item ($M = 75$, $SD = 13$) than the associative test ($M = 78$, $SD = 11$). The main effect of test length was not significant, $F(1, 39) = 3.03$, $MSE = .003$, $p = .090$, $\eta_p^2 = .07$, nor was the interaction, $F(1, 39) = 3.21$, $MSE = .01$, $p = .081$, $\eta_p^2 = .08$.[3]

The results of the present experiment show the predicted cross-over pattern of accuracy and confidence, which rules out a possibility that the confidence-accuracy dissociation obtained by Beaman et al. (2014) was a mere confound of the number of item presentations during test. We argue that this cross-over pattern in confidence and accuracy measures arises because when deciding about confidence in their recognition decisions, participants consider solely the absolute memory evidence supporting the alternative they endorsed as a target, ignoring evidence that would support the unchosen alternative. When associative and item recognition tests are contrasted, this absolute evidence for the chosen alternative is stronger in the associative than the item recognition tests. Targets endorsed in the associative recognition test are supported by both memory evidence indicating that individual items were studied and memory evidence indicating that a particular association linking two words was established at study. Note that while the use of item information within an associative recognition test may seem unintuitive because in this test all individual words were actually presented in the study phase, a large number of studies indicate that participants use this kind of non-diagnostic evidence in associative recognition tests (e.g., Buchler, Light, & Reder, 2008; Malmberg & Xu, 2007). By comparison, the chosen alternatives in the item recognition tests are supported mostly by evidence gathered for individual items. Evidence for associations is largely unavailable because these associations are not re-established at the time of a memory test (see Cohn & Moscovitch, 2007). Thus, on average there is more evidence supporting the chosen correct alternatives in the associative than in the item recognition test, resulting in higher confidence in this test, a pattern directly opposite to the pattern of recognition accuracy.

It is important to remain mindful that the cross-over dissociation between accuracy and confidence observed in Experiment 1 refers to a comparison of different memory tests for the same study materials. Even when confidence becomes dissociated from accuracy *across* testing conditions, it can still remain positively related *within* a test. To assess whether this is indeed the case, calibration curves for data pooled across participants separately for two types of test (collapsed across the test length factor) are presented in Fig. 2.

Visual inspection of the calibration curves reveals a clear positive relationship between confidence and accuracy on both tests. Thus, even though the use of different tests of memory dissociates confidence from accuracy *across* tests, confidence remains monotonically related to accuracy *within* a test. Note also that the calibration plots reveal that the curve for the item test is placed consistently higher than the curve for the associative test demonstrating that, at each level of accuracy, responses provided in the item test are given with lower confidence. This result indicates that the difference in confidence across tests is a general property of the data and is not caused solely by the most confidently held responses.

Experiment 1 revealed a striking dissociation between accuracy and confidence by which more accurate responding in the item recognition test was associated with less confidence. The robustness of this particular empirical pattern, however, has yet to be established. A potential criticism of our manipulation of the type of test is that the item recognition test – as implemented in our study – is somewhat unusual in that it requires endorsing as targets items that were not actually studied in the same form as presented in the test. While endorsing intact pairs as targets in the associative recognition test seems like a standard memory task, participants may become perplexed when asked to endorse particular combinations that were not actually studied – even though they comprise studied words – in the item recognition task. Being asked to do so could undermine their confidence in the decisions they make.

In Experiment 2 we sought a way of augmenting confidence by reassuring our participants that decisions they make in the item recognition task are by-and-large correct. To this end, we introduced feedback for recognition decisions in the procedure for Experiment 2. Previous studies indicate that feedback in recognition generally does not serve to improve accuracy (Kantner & Lindsay, 2010) so we do not predict any effects of this change in the procedure on the accuracy patterns. However, providing feedback across conditions differing in accuracy should boost confidence specifically in a condition in which accuracy is higher. With higher accuracy, participants should generally receive more correct feedback, which should then serve to augment confidence in recognition decisions they make. Here, interest centers on whether feedback would be sufficient to remove the confidence-accuracy

Table 1
Proportions of hits (accuracy) and means of confidence judgments for correct responses as a function of type of test (Experiments 1–3) and length of test (Experiments 1–2) conditions in Experiments 1–3. Standard deviations are provided in parentheses.

	Associative		Item		Item + association
	Short	Long	Short	Long	
Experiment 1					
Correct responses	.73 (.18)	.71 (.15)	.80 (.19)	.81 (.15)	–
Confidence in correct responses	80 (13)	77 (11)	74 (13)	75 (15)	–
Experiment 2					
Correct responses	.66 (.16)	.69 (.15)	.84 (.10)	.80 (.13)	–
Confidence in correct responses	74 (12)	74 (12)	73 (11)	77 (12)	–
Experiment 3					
Correct responses	.73 (.17)	–	.82 (.14)	–	.84 (.13)
Confidence in correct responses	85 (12)	–	81 (12)	–	84 (10)

[3] Given that the aim of this experiment was to confirm that the same pattern as the one found by Beaman et al. (2014) would emerge without the list length confound, we additionally conducted *t*-tests on the results from the short tests only. The same cross-over pattern of confidence and accuracy was found in this restricted (comprising one-third of all tested pairs) data set, with confidence in correct responses being lower in the item than the associative test, $t(39) = 2.73$, $p = .01$, $d = 0.43$, while for accuracy levels the trend was in the opposite direction, $t(39) = 1.86$, $p = .069$, $d = 0.30$.

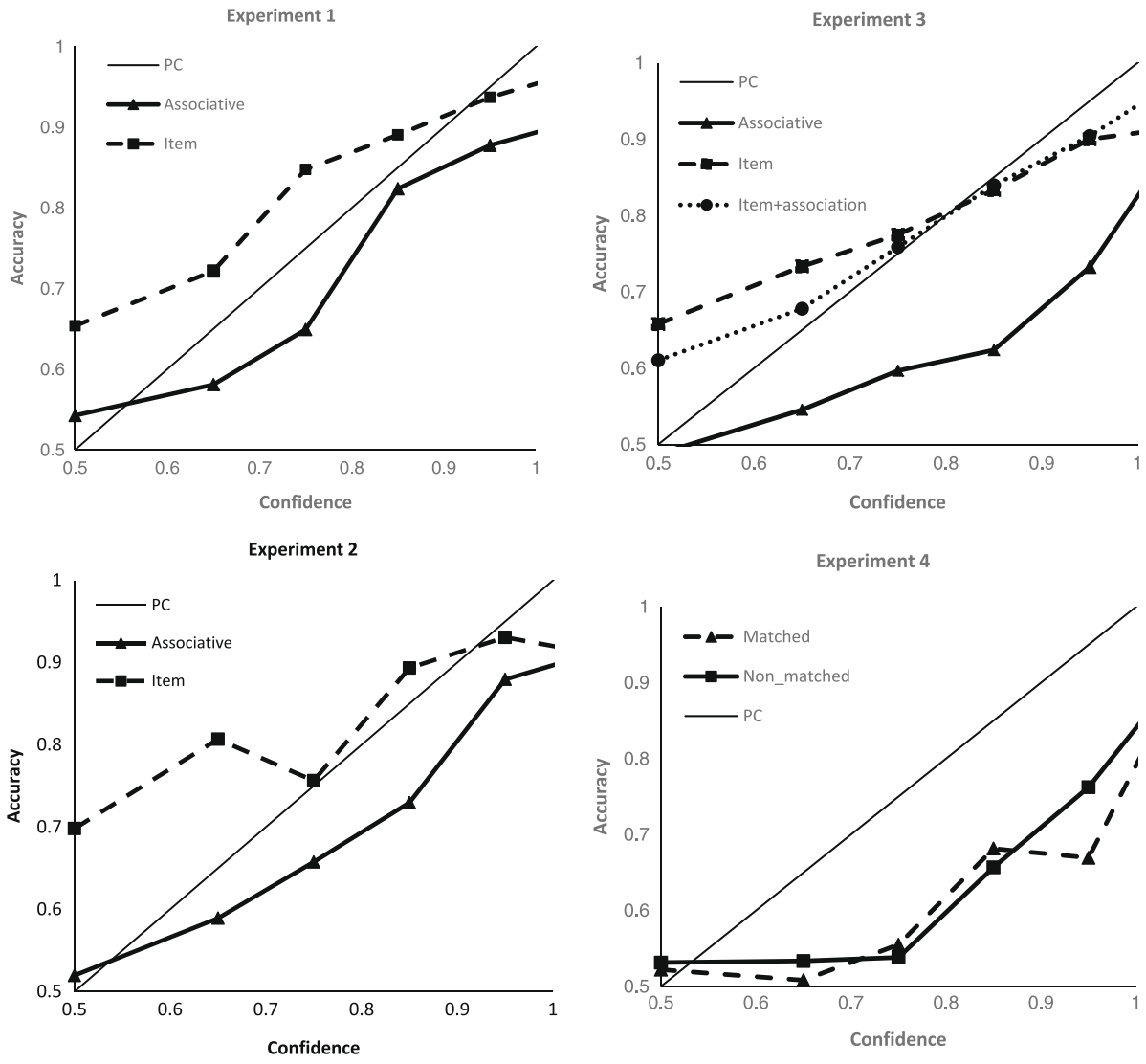

Fig. 2. Calibration curves in Experiments 1–4, based on the data collapsed across participants and for Experiments 1–2 also across the test length factor. Separate curves are plotted for different types of tests: associative and item recognition in Experiments 1 and 2, associative, item recognition and item + association in Experiment 3, and high vs. low choice similarity associative recognition in Experiment 4. Perfect calibration, denoted as PC, is provided for comparison.

dissociation revealed in Experiment 1. If not, the robustness of the confidence-accuracy dissociation would be underscored.

Method

Participants

Twenty-four Cardiff University students participated in exchange for course credit. The sample size was chosen to ensure the replication of the accuracy patterns observed in Experiment 1: with $\eta_p^2 = .17$ for this comparison, the required sample size to obtain power of .95 was 23.

Materials, design, and procedure

All materials were the same as in Experiment 1. We also used the same design, which varied both the type of test and the length of the study list. The results of Experiment 1 demonstrate that observed patterns of accuracy and confidence do not depend on the length of list and related repetition of words within the associative recognition test but the same design was here preserved in order to confirm this insight. The procedure was the same as in Experiment 1, except for the feedback that was presented after each response in a memory test (displayed for 1.5 s), informing participants whether their response was correct.

Results

The descriptive statistics are given in Table 1. Hit rates were subjected to a 2 (test type) × 2 (test length) repeated-measures ANOVA, which yielded significantly higher accuracy in the item recognition test (M = .82, SD = .10) than in the associative recognition test (M = .68, SD = .12), $F(1, 23) = 22.74$, $MSE = .02$, $p < .001$, $\eta_p^2 = .50$, but no main effect of test length, $F(1, 23) = 0.18$, $MSE = .01$, $p = .675$, $\eta_p^2 = .008$, and no interaction, $F(1, 23) = 2.29$, $MSE = .01$, $p = .144$, $\eta_p^2 = .09$. This parallels the results of Experiment 1. The ANOVA for confidence in correct responses yielded no significant effects: $F(1, 23) = 0.08$, $MSE = .007$, $p = .774$, $\eta_p^2 = .004$, for the main effect of test type, $F(1, 23) = 2.15$, $MSE = .004$, $p = .157$, $\eta_p^2 = .09$ for the main effect of test length, and $F(1, 23) = 1.55$, $MSE = .01$, $p = .226$, $\eta_p^2 = .06$ for the interaction.

A visual analysis of calibration curves (see Fig. 2) reveals two things. First, there is again a positive relationship between confidence and accuracy on both tests, pointing to a meaningful relationship between accuracy and confidence *within* each experimental condition. Second, the curve for the item recognition test is again consistently above the curve for the associative test, across all levels of accuracy.

Providing feedback after each memory response seemed to boost

participants' confidence to a higher level than in Experiment 1, to the point that there was no longer a significant difference in confidence between the associative and item recognition tasks. Nevertheless, differences in accuracy between the two tasks remained: a substantial difference – 13 percentage points – between item and associative recognition test accuracy was coupled to levels of confidence roughly comparable in magnitude for both tests.

The present results demonstrate that even with the provision of feedback pointing to highly accurate (81% correct) responding in the item recognition test, participants consistently undervalued the quality of their responses in this test, as compared to the associative recognition test. This is consistent with the absolute account of confidence in forced-choice recognition according to which confidence is based on evidence supporting the chosen alternative only, ignoring the evidence supporting unchosen alternatives. At the same time, as demonstrated by the calibration curves, confidence remains meaningfully related to accuracy within a given memory test.

Experiment 3

The aim of the first two experiments was to establish the veracity of the dissociation between accuracy and confidence first obtained by Beaman et al. (2014). The results presented so far are consistent in showing that while accuracy is higher when participants are asked to distinguish between intact and rearranged pairs than when they are asked to distinguish between rearranged and novel pairs, confidence does not track this difference. A particularly striking result was obtained in Experiment 1, where the confidence-accuracy dissociation took the shape of a cross-over pattern. This, we argue, was possible because by our test manipulation we varied simultaneously evidence supporting the chosen correct alternatives (targets) and the unchosen incorrect alternatives (lures). While variations in the former were responsible for differences in confidence, variations in the latter drove differences in accuracy, resulting in the cross-over pattern overall. In Experiment 3, we aimed to provide additional evidence for this hypothesis.

Arguably, the confidence-accuracy dissociation we expected to see has been thus far largely overlooked within the literature because memory studies tend to vary support for targets rather than lures, increasing confidence alongside accuracy. Here we argue that to reveal confidence-accuracy dissociations, evidence supporting lures – affecting accuracy but not confidence – needs to be varied across experimental conditions. Consequently, in Experiment 3 we designed conditions under which target strength was held constant while support for lures was varied. In Experiment 3, we included three types of recognition tests. We once again administered associative and item recognition tests, but this time we supplanted them with an item + association test in which participants were asked to distinguish between intact pairs and pairs consisting of two unfamiliar, not-previously-presented words (see Fig. 1). Because in this experiment feedback was not provided, we expected to replicate the cross-over pattern of accuracy and confidence that we first showed in our comparison of associative and item recognition tests in Experiment 1.

The crucial novel predictions concern the comparison between the associative and item + association test. These two tests hold targets constant, while varying only the nature of lures, with the associative test using rearranged pairs and the item + association test using novel pairs. With a difference in support for the lures, we predicted better recognition performance for the item + association test than for the associative test. We predict this on the basis that a difference in evidence supporting targets (equated across tests) and lures (larger for the associative recognition test) should favor correct responding in the item + association test. However, we also predicted that confidence would not track accuracy for this comparison. Specifically, assuming that confidence reflects solely support for the chosen alternative, we predicted that confidence in correct endorsements should be generally equated across these two tests that employ the same targets.

Notably, while a comparison of the associative and item + association tests allows for assessing confidence and accuracy across tests with the same targets but different lures, the present design allows also for the comparison of the item + association and item recognition tests, which has the potential to provide important insights. This comparison includes the same lures – novel pairs – but different targets, which are rearranged pairs for the item test and intact pairs for the item + association test. If confidence depends on the evidence supporting targets, the comparison of these two tests should yield a difference in confidence in correct responses, which should be higher for the item + association tests, when targets match memory records of both individual words and an association linking them.

The predictions made for the present experiment are not particularly novel inasmuch as the item + association test was also included in the study by Beaman et al. (2014). The patterns obtained there are – in hindsight – exactly as predicted by the absolute account of confidence in forced-choice recognition. While memory performance was higher in the item + association test (a difference of 10 percentage points), confidence (combined across correct and incorrect responses in this study) was almost exactly equal across these tests. Also, confidence was higher in the item + association test than in the item test. However, as noted earlier, the auditory distraction paradigm used by Beaman et al. was not designed explicitly to test the absolute evidence hypothesis. Here we provide a direct test of this hypothesis, eliminating all features of the design used by Beaman et al. that are not germane to the confidence-accuracy relationship, removing a confound between the type of test and item repetition within a test and focusing on confidence in correct responses, to which the predictions of the absolute account of confidence in forced-choice recognition directly apply.

Method

Participants

Thirty-two students and graduates of universities located in Warszawa and Łódź, Poland, were tested for monetary compensation. The sample size was increased in comparison to Experiment 2 to ensure good power for detecting differences in performance across tests, while also providing meaningful data with respect to the predicted null effect on confidence.

Materials, design, and procedure

Seven hundred and twenty Polish words were chosen from the frequency norms by Mandera, Keuleers, Wodniecka, and Brysbaert (2015). They were randomly paired to create 360 word-pairs, out of which 240 were divided into six study lists with 40 pairs per list and the remaining 120 pairs were divided into six sets of 20 to be used as lures in recognition tests. Participants studied six lists, each followed by a recognition test. There were three different types of test, each completed for two different study lists. The associative and item recognition tests were the same as in the respective short test conditions of Experiments 1 and 2. For the item + association test a randomly chosen half of study pairs were re-paired for each recognition trial to make novel pairs. Only short tests were used in the present experiment to avoid confounding the type of test and item repetitions across tests. The procedure for both the study phases and all recognition tests was the same as in Experiments 1 and 2, with studied pairs displayed for 1.5 s, with 500 ms interval, and unlimited time to provide both recognition decisions and following confidence judgments.

Results and discussion

The descriptive statistics are given in Table 1. Hit rates were subjected to a one-way repeated-measures ANOVA, which revealed significant differences across test conditions, $F(2, 62) = 8.96$, $MSE = .01$, $p < .001$, $\eta_p^2 = .224$. This reflected higher accuracy in both the item recognition test, $t(31) = 2.83$, $p = .008$, $d = 0.50$, and the item + association

recognition test, $t(31) = 3.81$, $p < .001$, $d = 0.67$, than in the associative recognition test, while performance did not significantly differ between item and item + association tests, $t(31) = 1.03$, $p = .31$, $d = 0.18$.

A one-way repeated-measures ANOVA on mean confidence judgments for correct responses also revealed reliable differences across conditions, $F(2, 62) = 5.18$, $MSE = 23.18$, $p = .008$, $\eta_p^2 = .143$. This reflected higher confidence in the associative recognition test, $t(31) = 2.93$, $p = .006$, $d = 0.52$, and in the item + association test, $t(31) = 2.34$, $p = .026$, $d = 0.41$, than in the item recognition test, while confidence in correct responses did not differ between associative and item + association tests, $t(31) = 0.765$, $p = .450$, $d = 0.13$.

A visual inspection of calibration curves (see Fig. 2) reveals two things. First, there is a positive relationship between confidence and accuracy on all types of tests, albeit less consistent for the associative and item + association tests at the highest levels of confidence. By and large, however, the plots again reveal meaningful relationships between accuracy and confidence *within* each test. Second, the curve for the associative recognition test is placed consistently lower than the plots for the two remaining test conditions across all levels of confidence showing a dissociation between confidence and accuracy *across* testing conditions. Once again, the calibration curves reveal that differences across testing conditions are not constrained to a subset of responses characterized by highest levels of confidence.

These results replicate Experiment 1 in showing a cross-over pattern of accuracy and confidence when associative and item recognition tests are compared. While accuracy is clearly higher in the item recognition test (by 9 percentage points), confidence in correct responses is actually higher in the associative recognition test. As argued throughout the present paper, these results can be explained if one assumes that while difference in evidence supporting two test alternatives underlies differences in accuracy across tests, confidence depends solely on evidence supporting the chosen alternative. The same absolute account of confidence predicts that confidence should become dissociated from accuracy when the associative recognition test is compared against the item + association test. While a difference in evidence supporting lures between these two tests should favor higher accuracy in the item + association test, the chosen correct alternative in these two tests is actually the same type of pair – intact – which should yield no difference in confidence if confidence does not take into account evidence supporting lures. This pattern was indeed observed. While accuracy was clearly higher in the item + association recognition test (by 11 percentage points), confidence in correct responses was almost identical across these tests, with – if anything – numerically higher confidence for correct responses in the associative recognition test. Because here we argue for a null result, we also report a Bayesian *t*-test with uninformed priors for this comparison, conducted using JASP software (JASP Team, 2019), which yielded a Bayes Factor of 4.04. This means that the data are 4.04 times more likely under the null than the alternative hypothesis, which can be considered moderate evidence (Lee & Wagenmakers, 2013) in favor of the null hypothesis.

One additional confidence-accuracy dissociation occurred in the present experiment. When item and item + association tests were compared, accuracy was roughly equated, yet confidence in correct responses was somewhat higher in the item + association test than in the item recognition test. The pattern of confidence here is again as predicted by the absolute account. Because intact pairs match memory records better than rearranged pairs, higher confidence should be observed when the former are endorsed. At the same time, however, the accuracy pattern for this comparison requires some comment. If there is more evidence for targets in the item + association test than in the item test, why is there no difference in accuracy? After all, increasing the evidence for targets should affect the balance of evidence between targets and lures, which determines accuracy.

There is a small numerical trend (2 percentage points) favoring accuracy in the item + association test, so this null effect could reflect insufficient power and thus the obtained confidence-accuracy dissociation may be more apparent than real[4]. Power could be insufficient for this particular comparison because the difference in balance of evidence caused by changing targets from rearranged to intact pairs is likely to be fairly small. In the present study, we manipulated this balance primarily by changing lures across tests. A difference in memory evidence between two previously studied words (for a rearranged pair) and two novel words should be large as it maps onto a difference between something that was studied and something that is completely new in the experimental context. By contrast, a difference in memory evidence between intact and rearranged pairs is likely to be much smaller because they both correspond to studied material. Such a small difference is unlikely to affect the balance of evidence substantially because it becomes diluted by noise coming from lures which are, on average, equated across conditions. By contrast, confidence – which depends solely on evidence supporting targets – may prove a much more sensitive measure of how well targets are remembered because confidence here is independent of evidence supporting lures (see Hanczakowski, Zawadzka, & Coote, 2014, for a similar argument). Hence, even if the strength of targets does affect balance of evidence, its influence may be much easier to detect in the measure of confidence, as it was in the present results.

To summarize, we have presented clear evidence that confidence in forced-choice recognition depends on the evidence supporting the chosen alternative, a mechanism that inevitably leads to confidence-accuracy dissociations because accuracy is determined by the balance of evidence between correct and incorrect alternatives. This novel evidence for the absolute account of confidence chimes with recent studies, using methods very different from the current ones, yet also providing consistent support for the absolute account of confidence in both perceptual (Zylberberg et al., 2012) and memory decisions (Miyoshi et al., 2018; Zawadzka et al., 2017). It extends these findings by showing that patterns predicted by the absolute account appear across tests varying the nature of support for correct responding: associative information in the associative recognition test, both item and associative information in the item + association test, and predominantly item information in the item recognition test. In all these tests confidence was based on evidence supporting the chosen alternative only, leading to confidence-accuracy dissociations across testing conditions. At the same time, it is vital to note another recent study devoted to the issue of confidence in forced-choice recognition, by Horry and Brewer (2016), in which the authors make a case for a relative account of confidence, where confidence depends on both evidence supporting targets and lures. We now turn to the discussion of their arguments, before presenting an experiment attempting to reconcile our theoretical stance with theirs.

Horry and Brewer (2016) shaped their experimental paradigm on the line-up procedure. Participants were presented with faces and each face was followed with a recognition trial in which participants were asked to select the just presented face from among a number of lures. The similarity of lures to the target face was varied. For example, in their Experiment 1 – using the simplest variant of the procedure – participants were presented with a target face accompanied by a single lure face which was either similar or dissimilar to the target. Subsequent experiments varied parameters such as the number of lures, whether the target was included in a recognition trial, and response options – allowing participants to 'reject' the trial if they thought the target face was not included. Independently of all these changes to the paradigm, a consistent pattern emerged such that increasing the similarity of lures to targets simultaneously decreased accuracy and reduced confidence in correct responses (when targets were endorsed). While the accuracy pattern is straightforwardly captured by the assumption that accuracy depends on the balance of evidence, the confidence pattern was

[4] Note that this difference was slightly larger (4 percentage points) and statistically significant in the study by Beaman et al. (2014).

interpreted by Horry and Brewer as consistent with the relative account of confidence, where confidence also depends on the balance of evidence. Under this argument, more similar lures reduce the relative difference in evidence supporting targets and lures, thus reducing confidence.

To understand why the results of Horry and Brewer (2016) do not necessarily provide support for the relative account of confidence, and can instead be accommodated by the absolute account, it is necessary to return to the study by Tulving (1981). So far, we have discussed a confidence-accuracy dissociation by which accuracy is higher in the A-X condition than in the A-B' condition (due to differences in balance of evidence) but this is not reflected in confidence in correct responses (due to equated evidence for targets on which confidence is based). However, as mentioned earlier, Tulving showed another pattern in this study by which accuracy was higher in the A-A' condition than the A-B' condition, yet confidence in correct responses showed an opposite pattern, being higher in the A-B' condition than the A-A' condition. This pattern of results led Tulving to distinguish between two types of similarity, which we will refer to as *memory similarity* and *choice similarity*. Memory similarity refers to how similar a recognition alternative is *to information stored in memory*. In a line-up example, this might be how similar a line-up member is to the memory representation of the actual culprit. Essentially, this is what we have referred to as memory evidence throughout the present manuscript. Thus, memory similarity is responsible for patterns observed for the A-B' and A-X conditions because it is the similarity of lures to memory representations that is varied across these pairs. Choice similarity refers to the similarity *between alternatives* included in a recognition trial. In the line-up example, if all members are bearded, or belong to the same ethnic group, then they are similar to each other on these dimensions regardless of how similar one or more members might also be to the culprit – who could be clean-shaven or a member of a different ethnic group[5]. The reinterpretation of results obtained by Tulving suggests that memory similarity of lures does not affect confidence – equated across A-B' and A-X conditions – but confidence is affected by choice similarity. Participants report higher confidence when recognition alternatives within a recognition trial are dissimilar to each other.

The distinction between memory and choice similarity allows the results obtained by Horry and Brewer (2016) to be reconciled with the absolute account of confidence. As Horry and Brewer acknowledge, their procedure does not distinguish between memory and choice similarity. Thus, in their low similarity condition, a lure was at the same time dissimilar from the target representation in memory and dissimilar from the target alternative in a given recognition trial. In this design, it is impossible to establish which type of similarity relationship drives differences in confidence. The role of memory similarity would be consistent with the relative account of confidence, but the role of choice similarity would mean that the results of Horry and Brewer (2016) may have no bearing on the distinction between absolute and relative accounts of confidence.

To argue that choice similarity can affect confidence requires more empirical support, which currently comes only from studies using the paradigm introduced by Tulving (1981). Here we attempted such a test using methods similar to the ones employed in Experiments 1–3. This served two aims. The first aim was to provide additional evidence for the idea that the patterns so far interpreted as supportive of the relative account of confidence (Horry & Brewer, 2016) may in fact reflect differences in choice similarity across experimental conditions. The second aim was to provide additional specificity to the absolute account of confidence. As discussed throughout this paper, the absolute account states that confidence in forced-choice recognition depends on the evidence supporting the chosen alternative, while ignoring the evidence supporting the unchosen alternative. This does not mean that unchosen alternatives are disregarded altogether, only that evidence supporting them as possible targets is not factored into confidence. Demonstrating that choice similarity affects confidence would thus serve to reveal how unchosen lures are capable of shaping confidence even though, as demonstrated throughout Experiments 1–3, their similarity to memory representations is irrelevant.

Experiment 4

Experiment 4 focused on the role of choice similarity—rather than memory similarity—in shaping confidence in forced-choice recognition using a setting comparing two variants of the associative recognition test. The first variant was the same as used in Experiments 1–3: on each recognition trial one intact pair and one recombined pair created from two words that were included in different studied pairs were pitted against each other. This was a condition of *low-choice similarity*. In the *high-choice similarity* condition, on each recognition trial an intact pair was pitted against a recombined pair which had the same first word as the intact pair but a different second word, taken from a different studied pair (see Fig. 1). A similar procedure was used in a study by Clark, Hori, and Callan (1993), who referred to the described conditions as NOLAP (no overlap across alternatives) and OLAP (overlap across alternatives). They found an accuracy advantage for forced-choice recognition for the NOLAP condition, but they did not include the confidence measure which is of main interest here.

In the present design, the *memory similarity* of studied pairs is equated across high-choice and low-choice similarity conditions: they both use only studied words and both contrast one intact and one recombined pair. However, these testing conditions differ in *choice similarity* insofar as the use of the same first words across test pairs creates higher choice similarity compared to a condition of no overlap. If choice similarity determines confidence in forced-choice recognition, we predict lower confidence in correct responses in the high-choice similarity condition—an equivalent of the A-A' condition in the study by Tulving (1981), as compared to the low-choice similarity condition—an equivalent of the A-B' condition.

Method

Participants

Forty-four students and graduates of universities located in Warszawa and Łódź, Poland, participated in return for monetary compensation. As we had no previous data from the same manipulation, we chose a sample size similar to that in our Experiment 1.

Materials, design, and procedure

A subset of 480 words from among Experiment 3 materials were chosen and assigned to six lists of 40 word pairs each. There were two types of tests: half the lists were followed by a low-choice similarity test, which was the same associative recognition test as in Experiment 3, and the other half of the lists were followed by a high-choice similarity recognition test, where each lure had the same first word as the intact pair used on the same test trial and the second word was chosen randomly from another pair, which did not serve as an intact pair in the test. The procedure for the present experiment was the same as in Experiment 3.

Results and discussion

A *t*-test comparing hit rates across the low- and high-choice similarity conditions failed to reveal a significant difference, $t(43) = 1.61$, $p =$

[5] In the literature concerning line-ups, members of a line-up other than a suspect are referred to as foils or fillers, whereas the term 'lure' is sometimes used in reference to an innocent suspect – a person included in a line-up who is not a culprit. However, for the sake of consistency, here we use the term 'lure' to refer to incorrect alternatives present in a particular recognition trials in all types of recognition tests, including line-ups.

.115, $d = 0.24$. If anything, accuracy was numerically higher in the low-choice similarity (M = .78, SD = .18) compared to the high-choice similarity condition (M = .76, SD = .16). This difference was reliable in the study by Clark et al. (1993), albeit in a slightly different design, employing a three-alternative forced choice test.

Of more importance, a *t*-test comparing confidence in correct responses revealed a significant difference, $t(43) = 2.03$, $p = .048$, $d = 0.31$, with higher confidence for the low-choice similarity condition ($M = 89$, $SD = 9$) than the high-choice similarity condition ($M = 87$, $SD = 9$). Because in the present experiment only associative recognition tests were administered and accuracy levels were roughly equated, this time we were able to also examine confidence in incorrect responses as most participants had data in all cells of the design. After excluding two participants with perfect performance in the low-choice similarity condition, there was a significant difference in confidence in incorrect responses, $t(41) = 3.04$, $p = .004$, $d = 0.47$, with higher confidence again for the low-choice similarity ($M = 72$, $SD = 2$) than the high-choice similarity condition ($M = 67$, $SD = 2$). The calibration curves are presented in Fig. 2. They reveal generally similar and positive curves for both conditions, which diverge only at the highest levels of confidence, where participants seem to become relatively less confident in the high-choice similarity condition than in the low-choice similarity condition.

The important conclusion from Experiment 4 is thus that choice similarity determines confidence in forced-choice recognition so that confidence is generally higher when alternatives within a given recognition trial are less similar to each other. It seems that when memory and choice similarity are co-varied – as they were in the study by Horry and Brewer (2016) – lower confidence for more similar lures cannot be taken as evidence that confidence judgments are a product of the balance of evidence between chosen and unchosen alternatives. Under these conditions, it is possible that confidence in correct responses is based only on evidence supporting targets, modulated by the similarity of alternatives on a given recognition trial. The present series of experiments shows that while the similarity of a lure to a memory representation of any of the studied items does not impact confidence (Experiments 1–3), the degree to which a lure is similar to the target on a given recognition trial does (Experiment 4). A residual issue relates to the reasons why confidence is higher in the low-choice similarity compared to the high-choice similarity condition. There are at least two possibilities.

First, Clark et al. (1993) obtained a parallel difference in accuracy—which was not reliable in our Experiment 4—and argued that this difference could be caused by increased recollection from using more studied items in the low choice similarity condition (the NOLAP condition in their study). By this reasoning, every single studied item may serve as a cue for a pair in which this word was embedded at study. Recollection of original associations for words used in recombined pairs can be used as a cue for recall-to-reject processes (Rotello & Heit, 2000), allowing for correct responding. With more individual items used in the low-choice similarity condition, there is thus greater opportunity for recollection and this additional recollection can also increase confidence. However, it is worth re-stating that the present experiment not only failed to replicate the accuracy advantage—which could be an issue of power that related to the Clark et al. study being better suited to revealing differences in accuracy (using three alternatives on each recognition trial) —but we also revealed, across experimental conditions, differences in confidence in incorrect responses. This difference is difficult to account for by evoking recall-to-reject processes which should generally give rise to correct responses only, as recollective processes are generally assumed to do (Yonelinas, 1994).

Another way of explaining reduced confidence in the high-choice similarity condition is to revisit the absolute account of confidence and consider it in light of the diagnostic feature-detection model of responding in line-ups developed by Wixted and Mickes (2014; see also Hanczakowski, Zawadzka, & Higham, 2014, for a similar approach to recognition tests). The diagnostic feature-detection model suggests that when all alternatives are presented together, not all of their features are treated with equal weight. Specifically, features shared across alternatives are discounted in favor of features that uniquely point towards one of the alternatives being a target. A similar mechanism of discounting non-diagnostic features, coupled with the absolute account of confidence, can be responsible for patterns of confidence discussed here. For the high-choice similarity condition, a non-diagnostic feature—a shared cue—can be discounted when arriving at a confidence judgment, thus reducing the overall strength of memory evidence on which confidence is based. Such discounting does not occur for the low-choice similarity trials, for which no features are shared across targets and lures. This account can explain the patterns for both correct and incorrect responses because the shared cue can be discounted as part of evidence supporting either the chosen target or the chosen lure. It has the advantage of explaining the whole pattern of confidence results obtained here while in theoretical terms having the distinction of linking successfully two accounts of responding in forced-choice recognition tests: the diagnostic feature-detection model and the absolute account of confidence.

General discussion

There is a common assumption— embedded in a number of models of recognition (Brown & Heathcote, 2008; Pleskac & Busemeyer, 2010; Ratcliff & Starns, 2009)— that when formulating a confidence judgment in recognition tests, what participants take into account is the balance of evidence favoring the chosen option over the unchosen option. In the case of forced-choice recognition, this means a balance between memory evidence supporting two recognition probes, whereas – in the case of an old/new recognition test – it could also mean a balance between evidence supporting two response options ("old" and "new") for a single probe. The greater the balance in favor of the chosen option, the higher the confidence attached to it. At the same time, in forced-choice recognition, this balance of evidence determines how accurate responding will be because both more evidence for the correct, studied alternative and less evidence for the incorrect, unstudied alternative point to the correct alternative. In this formulation both confidence and accuracy are shaped by the same factor, resulting in a strong relationship between confidence and accuracy across experimental conditions for the forced-choice recognition tests. In other words, changing accuracy by changing the balance of evidence means a corresponding change in confidence.

Evidence from the current series allows us to question these assumptions. It shows that while accuracy in forced-choice recognition does depend on the balance of evidence supporting two recognition probes, confidence displays a confirmation bias, being based on evidence supporting the chosen recognition probe only. While confirmation bias has been amply documented in other domains (Nickerson, 1998) it has been largely overlooked in the recognition memory literature. Only recently have studies begun to examine such formulations (Miyoshi et al., 2018; Zawadzka et al., 2017). As a result of this confirmation bias, confidence in forced-choice recognition may track accuracy within experimental conditions when differences in accuracy of different responses result from changes in evidence supporting *correct* alternatives, but confidence becomes dissociated from accuracy across experimental conditions when differences in accuracy result from changes in evidence supporting *incorrect* alternatives.

In this study, we manipulated the types of test by which memory for studied pairs was assessed. All tests were administered in the forced-choice format but they differed in what constituted a target and what constituted a lure. In Experiments 1, 2 and 3, we documented a confidence-accuracy dissociation by comparing associative and item recognition tests. These tests differed not only in terms of targets: intact pairs in the associative recognition test and rearranged pairs in the item recognition test, but also lures: rearranged pairs in the associative recognition test and novel pairs in the item recognition test. Changing both targets and lures across tests led to a cross-over pattern of confidence and accuracy, observed in Experiments 1 and 3: accuracy was

consistently higher in the item recognition test, while confidence in correct responses was higher in the associative recognition test. The confidence portion of this dissociation was eliminated in Experiment 2 when feedback concerning recognition decisions was provided, favoring high confidence in the condition producing higher accuracy. Differences in accuracy across test conditions remain unsurprising giving the nature of lures on these tests. Rearranged pairs constitute much stronger lures than novel pairs, which means that the balance of evidence should more consistently favor targets in the item recognition than the associative recognition test. What is surprising, however, is that confidence does not track this difference. We argue that this is because confidence in correct responses reflects not the balance of evidence but evidence supporting targets only and this is stronger for associative recognition than item recognition tests. This is because targets in the associative recognition test cue both item information and associative information linking these items, while targets in the item recognition test tend to cue only item information.

Additional support for the differential roles of target and lure strength in shaping accuracy and confidence comes from Experiment 3, in which another testing condition was included. In the item + association test, participants were asked to distinguish between intact and novel pairs. When compared to the associative recognition test, the item + association test has the same targets but different lures. Because accuracy depends on the balance of evidence for these two alternatives, there was a difference in accuracy between associative and item + association tests. However, because confidence does not depend on this balance, with equated targets there was no difference in confidence in correct responses across these tests. When compared to the item recognition test, the item + association test has the same lures but different targets. Here, there was no detectable difference in accuracy across these tests. This was likely because the balance of evidence is affected only slightly by relatively minor variations of strength of targets and becomes undetectable when noise associated with the lures is factored in. However, because confidence in correct responses is selectively attuned to the variations of memory evidence supporting targets, there was still a pattern of higher confidence for the item + association test, in which targets cued both item and associative information in memory, than for the item recognition test, in which targets cued predominantly item information.

The present results join a growing body of work showing that confidence in forced-choice recognition reflects absolute evidence supporting the chosen alternative. First inklings that confidence is focused on the evidence supporting the chosen alternative came from the literature on perception, where Zylberberg et al. (2012) showed such a pattern for judgments of movement direction. For some time the impact of this finding was not felt in the memory literature, focused as it is on manipulations introduced at encoding that vary the strength of targets rather than lures (but see Jou et al., 2016, for a relevant discussion). Zawadzka et al. (2017) overcame this problem with the use of the plurals paradigm, in which evidence for lures can be varied by varying the strength of the parent word and showed that confidence reflects the strength of the chosen alternative, while being independent of the strength of the unchosen alternative. Because Zawadzka et al. used a procedure with almost chance levels of performance, they were able to analyze meaningfully incorrect responses and found the same pattern: when a lure was chosen, confidence in this choice reflected the strength of the chosen lure, not the balance of evidence between the lure and the target on a given trial. A similar approach was recently adopted by Miyoshi et al. (2018), who varied the memorability of target pictures and focused on confidence in incorrect choices. The balance of evidence would predict that with growing memorability of targets, confidence in incorrect responses should be reduced. However, an opposite pattern was actually documented by Miyoshi et al., who showed that with increasing strength of targets, confidence in incorrect responses also increased.

The pattern where increased strength of the unchosen alternative leads to increased rather than decreased confidence in the correctness of the chosen alternative is the exact opposite of what the balance of evidence account of confidence would predict. However, is it what the absolute account would always predict? It is worth noting that such a pattern was not observed in the present study. In Experiment 3, a comparison between associative and item + association recognition tests held targets equal and varied the strength of lures, with this factor having no effect on confidence when targets were endorsed. By the absolute account of confidence, the greater strength of lures in the associative recognition test could make endorsed targets particularly strong, producing greater confidence in correct responses than for the item + association recognition test – the 'skimming off' of responses by strong but unchosen alternatives (Miyoshi et al., 2018). However, this pattern of 'skimming off' is not universal as it was not observed for the present Experiment 3, nor was it observed for the comparison of A – B' and A – X pairs in the study by Tulving (1981). We suggest that these different outcomes reflect differences in the proportions of responses affected by the 'skimming off' by unchosen alternatives. Miyoshi et al. focused on incorrect responses, which are much less in evidence than correct responses in the most of recognition tests. The strength of any unchosen correct alternative has a large effect on the population of chosen—but incorrect—alternatives. Only the very strongest incorrect alternatives are endorsed when particularly strong correct items are present on the same recognition trial. By contrast, both Tulving (1981) in his study on picture recognition and we in the present Experiment 3 focused on correct responses only. Changing the strength of the unchosen-incorrect alternative across tests has necessarily a proportionally much smaller effect on the population of the chosen-correct alternatives. Whatever the difference in strength of lures, the majority of targets are endorsed anyway because on average they match memory better than either weak or strong lures. Thus, the fact that the majority of endorsed correct alternatives remain the same across conditions differing in the strength of unchosen incorrect alternatives is likely to undermine the possibility of detecting the 'skimming off' effects. This question awaits further research, but we note here that both the 'skimming off' pattern and the situation in which confidence remains entirely unaffected by the strength of the unchosen alternative remain clearly inconsistent with the balance-of-evidence account of confidence.

An important question that has not been addressed in the present study is why people would choose to disregard information contained in the unchosen alternative when making their confidence judgments. A hypothesis concerning the reason for the adoption of the absolute heuristic for formulating confidence judgments in forced-choice test has been recently proposed by Miyoshi and Lau (2020). Using the Signal Detection Theory (SDT) approach to forced-choice recognition, Miyoshi and Lau showed by means of a series of simulations how metacognitive resolution – people's ability to discriminate between their own correct and incorrect responses with their confidence judgments – depends on basic memory ability. Surprisingly, these simulations showed that there are conditions under which the absolute heuristic for formulating confidence judgments – what Miyoshi and Lau referred to as basing confidence on response-congruent evidence – leads to better metacognitive resolution of these judgments than using the balance-of-evidence approach. The conditions that favored the absolute heuristic were related to a ratio of variances associated with distributions of targets and lures. By increasing the ratio of target distribution variance to lure distribution variance, the benefits of the absolute heuristic outweigh the balance-of-evidence heuristic. Importantly, studies on recognition memory generally show that the standard deviation of the target distribution is indeed larger than the standard deviation of lure distribution, with the usual value of this ratio of approximately 1.25 (Mickes, Wixted, & Wais, 2007). The fact that in recognition memory tasks evidence for targets is more variable than evidence for lures, and these are precisely the conditions which lead to benefits for metacognitive resolution when confidence judgments are based on absolute evidence, suggests that the patterns observed here may be grounded in

participants' drive towards metacognitive optimality. If using a certain strategy consistently leads to metamemory benefits, then adopting this strategy also in experimental conditions may be a result of life-long training.

One point needs to be underscored regarding the issue of variances associated with evidence supporting targets and lures in forced-choice recognition tests – a crucial element contributing to potential benefits for metacognitive resolution of applying the absolute heuristic according to the simulations presented by Miyoshi and Lau (2020). While this variance is generally higher for targets than for lures in single-item recognition of the type examined by Zawadzka et al. (2017) and Miyoshi et al. (2018), this difference seems particularly pronounced if targets, but not lures, are characterized by strong associative information. Kelley and Wixted (2001) considered the variability of item and associative information in different variants of associative recognition tests. They established that the ratio of standard deviations of intact pairs and novel pairs in these tests is 2.0, compared to 1.25 observed for single item targets and lures. They argued that this higher ratio of variances is generally due to a greater variability of associative compared to item information, with associative information contributing to detection of intact pairs but not rejection of lures in the form of novel pairs. Following the logic of Miyoshi and Lau, a test pitting against each other intact and novel pairs should thus be one that leads to a particularly strong benefit of using the absolute heuristic for the metacognitive resolution of confidence judgments. The fact that people indeed seem to use the absolute heuristic when providing confidence judgments in the item + association recognition test is consistent with the idea that this heuristic is used under conditions that are likely to promote greater metacognitive resolution than one that could be possibly achieved by applying the balance-of-evidence approach.

At the same time, it is important to note that in order to account for the results of our study, it is necessary to assume that the same heuristic was used across all types of recognition tests. Only then can one fully describe the pattern obtained in Experiment 3, where tests with negligible differences in terms of performance – item and item + association recognition – displayed differences in terms of confidence, while tests characterized by large differences in terms of performance – associative and item + association recognition – showed virtually no differences in terms of confidence. Kelley and Wixted's (2001) observations regarding variability of associative information suggest that tests considered here must also have differed vastly in terms of variability across targets and lures. Cohn and Moscovitch (2007) considered the same types of tests as the ones employed in the present study and demonstrated that associative information is likely to contribute both to evidence supporting targets (recall-to-accept) and evidence discrediting lures (recall-to-reject) in the associative recognition test, is likely to contribute via both automatic and strategic processing to evidence supporting targets in the item + association test, and is much less likely to contribute to evidence supporting targets in the item recognition test due to a disruption to automatic processing that does not transpire for rearranged words. All in all, the contribution of associative information to memory performance seems thus to diminish progressively from associative recognition, through item + association recognition, to item recognition. The fact that the absolute heuristic for formulating confidence judgments seemed to be used across all these tests indicates that the use of this strategy is not scaled to match the potential benefits for metacognitive resolution, which – as suggested by Miyoshi and Lau (2020) – should emerge mostly when associative information contributes much to target recognition but not to lure rejection – conditions most alike to those in the item + association recognition test.

This persistence of the absolute heuristic is also consistent with the results of Zawadzka et al. (2017), who showed that it explains the patterns of confidence across recognition trials in a forced-choice recognition test for single items in which variances of targets and lures were kept constant (because additional repetitions of targets did not lead to increased variance, see Starns, & Ratcliff, 2014). Taken together, the work by Miyoshi and Lau (2020) indicates the specific conditions under which the absolute heuristic is likely to produce superior metacognitive resolution, and thus goes a long way towards explaining why people may be not motivated to factor evidence supporting the unchosen alternative into their confidence judgments. However, it seems that the absolute heuristic might well be a very general strategy for rendering confidence judgments, observed across tests for various stimuli like single words (Jou et al., 2016; Zawadzka et al., 2017), pictures (Miyoshi et al., 2018; Tulving, 1981), pairs of words, and consequently also in tests in which stimuli may be characterized by various levels of contributions of item and associative information. The issue of boundary conditions – if any – of using this heuristic for confidence judgments needs to be the topic of further research.

We started the present paper with examples from the area of eyewitness memory, where – in the case of line-ups – the role of confidence judgments has been underscored in recent years (see Wixted et al., 2015). As highlighted earlier, there is now a growing consensus that confidence judgments should be elicited when testimonies and identifications are made because such judgments tend to be informative, being generally related to accuracy of memory responses, at least when memory is tested for the first time (Wixted & Wells, 2017). Although the present work has been concerned with somewhat artificial versions of forced-choice recognition procedures, seemingly far removed from the case of line-ups, its results may have some bearing on the understanding of confidence judgments in this more applied setting. First, across our four experiments we plotted calibration curves which revealed that indeed more accurate responding seemed to be related to greater confidence within each of the recognition test we administered. This general result confirms that confidence is informative with regards to accuracy. At the same time, the patterns of cross-conditions dissociations do not remain without consequences for anyone who is interested in interpreting another person's confidence judgments. Looking again at calibration curves, it is important to note that while confidence remains predictive of accuracy within a given test, the tests themselves differ widely in terms of average levels of confidence quite independently of differences in accuracy. While people may be quite accurate when responding based on familiarity of individual words, as in the item recognition tests examined here, they are unlikely to be particularly confident in these decisions. This observation joins previous reports showing that familiarity-based decisions are generally characterized by low confidence, whether rendered with regard to faces (Reinitz, Séguin, Peria, & Loftus, 2012) or pictures (Reinitz et al., 2011). Also, while people may be relatively inaccurate when responding in the face of misleading lures, as in the associative recognition tests, they are likely to be confident in a quite unwarranted way. This highlights the importance of considering the context in which recognition decisions are made for understanding the meaning of confidence judgments. A misleading context in the form of similar lures is very likely to undermine the accuracy of recognition responses, but this is unlikely to find its reflection in confidence judgments, making them less than reliable indicator of accuracy.

This last argument seems at first blush particularly relevant to the issue of line-up identifications in which the question of similarity of lures—whether a line-up can be called fair—received much attention from researchers (e.g., Wells, Small, Penrod, Malpass, Fulero, & Brimacombe, 1998). Indeed, we have discussed at length the study by Horry and Brewer (2016), who manipulated lure similarity across recognition procedures that were built largely based on the structure of line-ups and showed that greater lure similarity undermined confidence when targets were endorsed. We argue that memory evidence supporting lures does not impact upon confidence, but their similarity to the target present in the particular recognition trial – choice similarity – does affect confidence, with greater similarity leading to lower confidence, as shown in our Experiment 3. However, our recognition procedure differs from the line-up procedure in two important points. First, it does not include either trials for which targets would be absent –

similar to line-ups including an innocent suspect but not the culprit – nor does it provide an option to respond 'neither', which would be similar to rejecting a line-up. It is worth noting that a recent model-fitting study, using an SDT approach to line-ups, revealed that confidence in line-up decisions is well described by the balance-of-evidence approach (Wixted, Vul, Mickes, & Wilson, 2018). It is possible that a crucial difference underlying our results with recognition tests on the one hand and line-ups on the other is based on the availability of different response options in these tasks. This hypothesis could be tested in future studies employing recognition tests built to be similar to line-up procedures and thus including options to 'reject' recognition trials. Such a variant of recognition testing was recently developed by Finley, Wixted, and Roediger (2020) and can be used to bring closer the currently discrepant studies on metacognitive judgments in the recognition and line-up domains.

Another way in which our recognition tasks differ from line-ups is that whereas we presented our participants with a large number of recognition trials, eyewitnesses are usually provided only with a single line-up. When there are multiple recognition trials, it is possible to at least partially disentangle choice similarity and similarity between probes and related memory representations. On a given trial, a lure can be similar to the target on the same trial, similar to a different target, or dissimilar to all targets, very much like in the procedure used by Tulving (1981). In a line-up task, a lure that is similar to the memory representation of the target is also similar to the target – or its substitute in a target-absent line-up – presented in the given line-up because there is actually only a single target. Thus, in line-ups, memory and choice similarity are almost necessarily confounded in the same way they were confounded in the study by Horry and Brewer (2016). Under these circumstances, the distinction between choice and memory similarity becomes irrelevant because on any account similar lures should undermine confidence, as Horry and Brewer demonstrated. Our study provides important novel insights regarding the basis of confidence in forced-choice recognition but also highlights the need to consider the particulars of a given recognition task to make inferences regarding the role of lures in shaping confidence.

CRediT authorship contribution statement

Maciej Hanczakowski: Conceptualization, Methodology, Software, Formal analysis, Investigation, Resources, Writing - original draft, Supervision, Project administration, Funding acquisition. **Ewa Butowska:** Methodology, Investigation, Resources, Writing - review & editing. **C. Philip Beaman:** Conceptualization, Writing - review & editing. **Dylan M. Jones:** Conceptualization, Writing - review & editing. **Katarzyna Zawadzka:** Conceptualization, Methodology, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Acknowledgement

This research was partially supported by grant 2017/27/B/HS6/02001 from the National Science Centre in Poland awarded to Maciej Hanczakowski, and by grant PPN/PPO/2018/1/00103 from the Polish National Agency for Academic Exchange awarded to Katarzyna Zawadzka.

All data are available online at https://osf.io/yb867/.

References

Beaman, C. P., Hanczakowski, M., & Jones, D. M. (2014). The effects of distraction on metacognition and metacognition on distraction: Evidence from recognition memory. *Frontiers in Psychology, 5*, 439. https://doi.org/10.3389/fpsyg.2014.00439.

Besken, M., & Mulligan, N. W. (2013). Easily perceived, easily remembered? Perceptual interference produces a double dissociation between metamemory and memory performance. *Memory & Cognition, 41*, 897–903. https://doi.org/10.3758/s13421-013-0307-8.

Brewer, N., & Wells, G. L. (2016). The confidence-accuracy relationship in eyewitness identification: Effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied, 12*, 11–30. https://doi.org/10.1037/1076-898x.12.1.11.

Brewer, W. F., & Sampaio, C. (2006). Processes leading to confidence and accuracy in sentence recognition: A metamemory approach. *Memory, 14*, 540–552. https://doi.org/10.1080/09658210600590302.

Bröder, A., & Schütz, J. (2009). Recognition ROCs are curvilinear – or are they? On premature arguments against the two-high-threshold model of recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 587–606. https://doi.org/10.1037/a0015279.

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology, 57*, 153–178. https://doi.org/10.1016/j.cogpsych.2007.12.002.

Buchler, N. G., Light, L. L., & Reder, L. M. (2008). Memory for items and associations: Distinct representations and processes in associative recognition. *Journal of Memory and Language, 59*, 183–199. https://doi.org/10.1016/j.jml.2008.04.001.

Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic Bulletin & Review, 7*, 26–48. https://doi.org/10.3758/bf03210724.

Chandler, C. C. (1994). Studying related pictures can reduce accuracy, but increase confidence, in a modified recognition test. *Memory & Cognition, 22*, 273–280. https://doi.org/10.3758/bf03200854.

Clark, S. E. (1997). A familiarity-based account of confidence-accuracy inversions in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 232–238. https://doi.org/10.1037/0278-7393.23.1.232.

Clark, S. E., Hori, A., & Callan, D. E. (1993). Forced-choice associative recognition: Implications for global-memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 871–881. https://doi.org/10.1037/0278-7393.19.4.871.

Cohn, M., & Moscovitch, M. (2007). Dissociating measures of associative memory: Evidence and theoretical implications. *Journal of Memory and Language, 57*, 437–454. https://doi.org/10.1016/j.jml.2007.06.006.

Dobbins, I. G., Kroll, N. E. A., & Liu, Q. (1998). Confidence-accuracy inversions in scene recognition: A remember-know analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 1306–1315. https://doi.org/10.1037/e536982012-393.

Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOLs) and the delayed-JOL effect. *Memory & Cognition, 20*, 373–380. https://doi.org/10.3758/bf03210921.

Finley, J. R., Wixted, J. T., & Roediger, H. L. (2020). Identifying the guilty word: Simultaneous versus sequential lineups for DRM word lists. *Memory & Cognition, 48*, 903–919. https://doi.org/10.3758/s13421-020-01032-6.

Grabman, J. H., Dobolyi, D. G., Berelovich, N. L., & Dodson, C. S. (2019). Predicting high confidence errors in eyewitness memory: The role of face recognition ability, decision-time, and justifications. *Journal of Applied Research in Memory and Cognition, 8*, 233–243. https://doi.org/10.1016/j.jarmac.2019.02.002.

Hanczakowski, M., Zawadzka, K., & Coote, L. (2014). Context reinstatement in recognition: Memory and beyond. *Journal of Memory and Language, 72*, 85–97. https://doi.org/10.1016/j.jml.2014.01.001.

Hanczakowski, M., Zawadzka, K., & Higham, P. A. (2014). The dud-alternative effect in memory for associations: Putting confidence into local context. *Psychonomic Bulletin & Review, 21*, 543–548. https://doi.org/10.3758/s13423-013-0497-x.

Hanczakowski, M., Pasek, T., Zawadzka, K., & Mazzoni, G. (2013). Cue familiarity and 'don't know' responding in episodic memory tasks. *Journal of Memory and Language, 69*, 368–383. https://doi.org/10.1016/j.jml.2013.04.005.

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology, 56*, 208–216. https://doi.org/10.1037/h0022263.

Higham, P. A. (2002). Strong cues are not necessarily weak: Thomson and Tulving (1970) and the encoding specificity principle revisited. *Memory & Cognition, 30*, 67–80. https://doi.org/10.3758/bf03195266.

Higham, P. A. (2007). No special K! A signal detection framework for the strategic regulation of memory accuracy. *Journal of Experimental Psychology: General, 136*, 1–22. https://doi.org/10.1037/0096-3445.136.1.1.

Heathcote, A., Bora, B., & Freeman, E. (2010). Recollection and confidence in two-alternative forced-choice episodic recognition. *Journal of Memory and Language, 62*, 183–203. https://doi.org/10.1016/j.jml.2009.11.003.

Hintzman, D. L., Curran, T., & Oppy, B. (1992). Effects of similarity and repetition on memory: Registration without learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*, 667–680. https://doi.org/10.1037/0278-7393.18.4.667.

Horry, R., & Brewer, N. (2016). How target-lure similarity shapes confidence judgments in multiple-alternative decision tasks. *Journal of Experimental Psychology: General, 145*, 1615–1634. https://doi.org/10.1037/xge0000227.

JASP Team (2019). JASP (Version 0.9) [Computer software]. Retrieved from https://jasp-stats.org/.

Jou, J., Flores, S., Cortes, H. M., & Leka, B. G. (2016). The effects of weak versus strong relational judgments on response bias in two-alternative- forced-choice recognition: Is the test criterion-free? *Acta Psychologica, 167*, 30–44. https://doi.org/10.1016/j.actpsy.2016.03.014.

Kantner, J., & Lindsay, D. S. (2010). Can corrective feedback improve recognition memory? *Memory & Cognition, 38*, 389–406. https://doi.org/10.3758/mc.38.4.389.

Kelley, C. M., & Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language, 32*, 1–24. https://doi.org/10.1006/jmla.1993.1001.

Kelley, R., & Wixted, J. T. (2001). On the nature of associative information in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 701–722. https://doi.org/10.1037/0278-7393.27.3.701.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*, 349–370. https://doi.org/10.1037/0096-3445.126.4.349.

Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological Review, 119*, 80–113. https://doi.org/10.1037/a0025648.
Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103*, 490–517. https://doi.org/10.1037/0033-295x.103.3.490.
Lee M. D., & Wagenmakers E. J. (2013). Bayesian cognitive modeling: A practical course. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/cbo9781139087759.
Malmberg, K. J., & Xu, J. (2007). On the flexibility and the fallibility of associative memory. *Memory & Cognition, 35*, 545–556. https://doi.org/10.3758/bf03193293.
Mandera, P., Keuleers, E., Wodniecka, Z., & Brysbaert, M. (2015). Subtlex-pl: Subtitle-based word frequency estimates for Polish. *Behavioral Research Methods, 47*, 471–483. https://doi.org/10.3758/s13428-014-0489-4.
Mazancieux, A., Fleming, S. M., Souchay, C., & Moulin, C. J. A. (2020). Is there a g factor for metacognition? Correlations in retrospective metacognitive sensitivity across tasks. *Journal of Experimental Psychology: General, 149*, 1788–1799. https://doi.org/10.1037/xge0000746.
Mickes, L., Wixted, J. T., & Wais, P. E. (2007). A direct test of the unequal-variance signal detection model of recognition memory. *Psychonomic Bulletin & Review, 14*, 858–865. https://doi.org/10.3758/bf03194112.
Miyoshi, K., Kuwahara, A., & Kawaguchi, J. (2018). Comparing the confidence calculation rules for forced-choice recognition memory: A winner-takes-all rule wins. *Journal of Memory and Language, 102*, 142–154. https://doi.org/10.1016/j.jml.2018.06.001.
Miyoshi, K., & Lau, H. (2020). A decision-congruent heuristic gives superior metacognitive sensitivity under realistic variance assumptions. *Psychological Review, 127*, 655–671. https://doi.org/10.1037/rev0000184.
Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*, 175–220. https://doi.org/10.1037/1089-2680.2.2.175.
Palmer, M., Brewer, N., Weber, N., & Nagesh, A. (2013). The confidence-accuracy relationship for eyewitness identification decisions: Effects of exposure duration, retention interval, and divided attention. *Journal of Experimental Psychology: Applied, 19*, 55–71. https://doi.org/10.1037/a0031602.
Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review, 117*, 864–901. https://doi.org/10.1037/a0019737.
Ratcliff, R., & Starns, J. J. (2009). Modelling confidence and response time in recognition memory. *Psychological Review, 116*, 59–83. https://doi.org/10.1037/a0014086.
Reinitz, M. T., Peria, W. J., Séguin, J. A., & Loftus, G. R. (2011). Different confidence-accuracy relationships for feature-based and familiarity-based memories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 507–515. https://doi.org/10.1037/a0021961.
Reinitz, M. T., Séguin, J. A., Peria, W. J., & Loftus, G. R. (2012). Confidence-accuracy relations for faces and scenes: Roles of features and familiarity. *Psychonomic Bulletin & Review, 19*, 1085–1093. https://doi.org/10.3758/s13423-012-0308-9.
Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology: General, 137*, 615–625. https://doi.org/10.1037/a0013684.
Roediger, H. L., Wixted, J. T., & DeSoto, K. A. (2012). The curious complexity between confidence and accuracy in reports from memory. In L. Nadel, & W. Sinnott-Armstrong (Eds.), *Memory and law* (pp. 84–118). Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199920754.003.0004.
Rotello, C. M., & Heit, E. (2000). Associative recognition: A case of recall-to-reject processing. *Memory & Cognition, 28*, 907–922. https://doi.org/10.3758/bf03209339.
Semmler, C., Brewer, N., & Wells, G. L. (2004). Effects of postidentification feedback on eyewitness identification and nonidentification confidence. *Journal of Applied Psychology, 89*, 334–346. https://doi.org/10.1037/0021-9010.89.2.334.
Starns, J. J., Chen, T., & Staub, A. (2017). Eye-movements in forced-choice recognition: Absolute judgments can preclude relative judgments. *Journal of Memory and Language, 93*, 55–66. https://doi.org/10.1016/j.jml.2016.09.001.
Starns, J. J., & Ksander, J. C. (2016). Item strength influences source confidence and alters source memory zROC slopes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*, 351–365. https://doi.org/10.1037/xlm0000177.
Starns, J. J., & Ratcliff, R. (2014). Validating the unequal-variance assumption in recognition memory using response time distributions instead of ROC functions: A diffusion model analysis. *Journal of Memory and Language, 70*, 36–52. https://doi.org/10.1016/j.jml.2013.09.005.
Tulving, E. (1981). Similarity relations in recognition. *Journal of Verbal Learning and Verbal Behavior, 20*, 479–496. https://doi.org/10.1016/s0022-5371(81)90129-8.
Undorf, M., Zimdahl, M. F., & Bernstein, D. M. (2017). Perceptual fluency contributes to effects of stimulus size on judgments of learning. *Journal of Memory and Language, 92*, 293–304. https://doi.org/10.1016/j.jml.2016.07.003.
Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. A. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human Behavior, 22*, 603–647. https://doi.org/10.1023/a:1025750605807.
Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114*, 152–176. https://doi.org/10.1037/0033-295x.114.1.152.
Wixted, J. T., & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review, 121*, 262–276. https://doi.org/10.1037/a0035940.
Wixted, J. T., Mickes, L., Clark, S. E., Gronlund, S. D., & Roediger, H. L. (2015). Initial eyewitness confidence reliably predicts eyewitness identification accuracy. *American Psychologist, 70*, 515–526. https://doi.org/10.1037/a0039510.
Wixted, J. T., Vul, E., Mickes, L., & Wilson, B. W. (2018). Models of lineup memory. *Cognitive Psychology, 105*, 81–114. https://doi.org/10.1016/j.cogpsych.2018.06.001.
Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest, 18*, 10–65. https://doi.org/10.1177/1529100616686966.
Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence from a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 1341–1354. https://doi.org/10.1037/0278-7393.20.6.1341.
Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*, 441–517. https://doi.org/10.1006/jmla.2002.2864.
Zawadzka, K., & Higham, P. A. (2016). Recalibration effects in judgments of learning: A signal detection analysis. *Journal of Memory and Language, 90*, 161–176. https://doi.org/10.1016/j.jml.2016.04.005.
Zawadzka, K., Higham, P. A., & Hanczakowski, M. (2017). Confidence in forced-choice recognition: What underlies the ratings? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*, 552–564. https://doi.org/10.1037/xlm0000321.
Zylberberg, A., Barttfeld, P., & Sigman, M. (2012). The construction of confidence in a perceptual decision. *Frontiers in Integrative Neuroscience, 6*, 79. https://doi.org/10.3389/fnint.2012.00079.